PATENT ABSTRACTS OF JAPAN

(11)Publication number : 11-345239

(43)Date of publication of application : 14.12.1999

(51)Int.Cl. G06F 17/30

(21)Application number : 10-153231

(22)Date of filing : 02.06.1998

(71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>

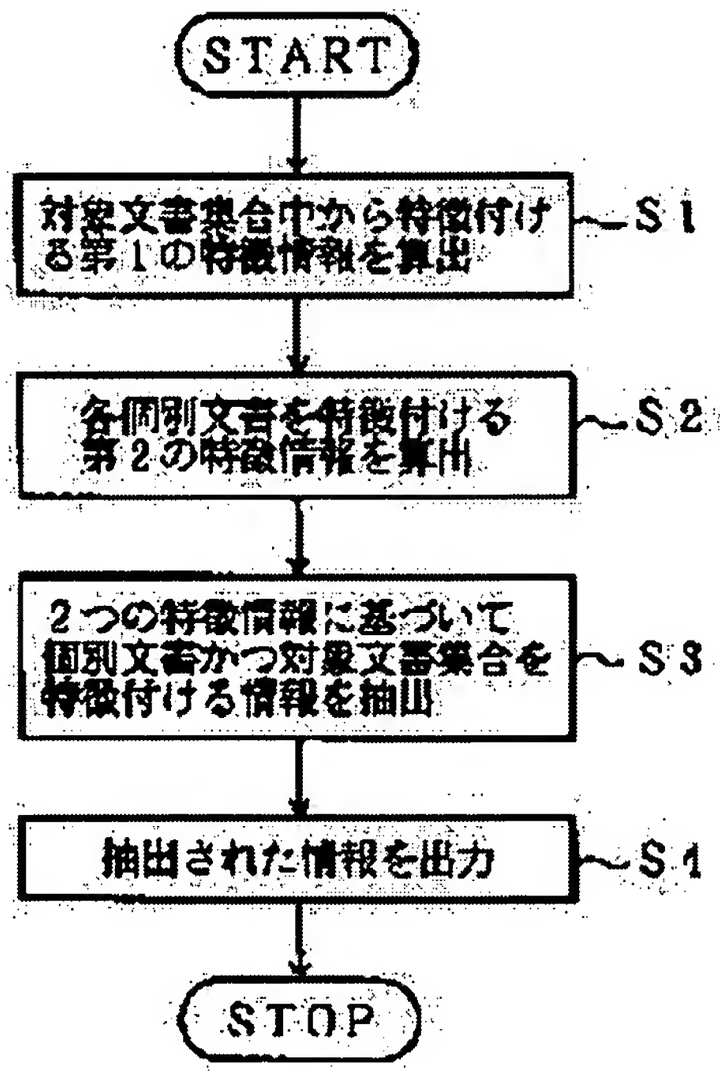
(72)Inventor : IZUDERA TOSHINORI

(54) METHOD AND DEVICE FOR EXTRACTING DOCUMENT INFORMATION, AND STORAGE MEDIUM STORED WITH DOCUMENT INFORMATION EXTRACTION PROGRAM

(57)Abstract:

PROBLEM TO BE SOLVED: To provide the method and device for extracting a document by which even new information can be extracted without preliminary preparing a pattern, heuristics, etc., and to provide a storage medium in which a document information extracting program is stored.

SOLUTION: The 1st characteristic information which characterizes an object document set with respect to a standard document set is calculated from the object document set (S1), the 2nd characteristic information which characterizes each separate document in the object document set with respect to other separate documents is calculated from each separate document in the object document set (S2), the information which characterizes the object document set more and also characterizes each separate document with respect to the other separate documents is extracted from the each separate document (S3) based on the 1st and 2nd characteristic information, and the extracted information is outputted as the information that characterizes each separate document.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報**（Ａ）　(11)特許出願公開番号

特開平11－345239

(43)公開日　平成11年(1999)12月14日

(51)Int.Cl.[6]	識別記号	ＦＩ		
Ｇ０６Ｆ　17/30		Ｇ０６Ｆ	15/401	３１０Ａ
			15/40	３７０Ａ

審査請求　未請求　請求項の数11　ＯＬ　（全 20 頁）

(21)出願番号　特願平10－153231

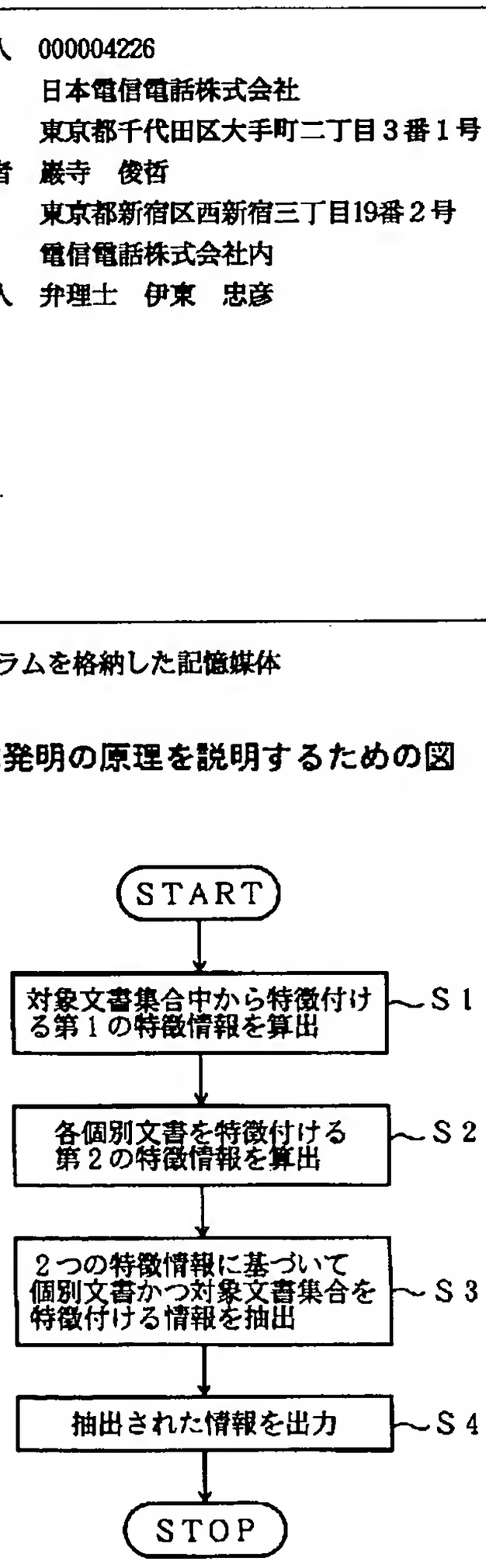
(22)出願日　平成10年(1998)６月２日

(71)出願人　000004226
日本電信電話株式会社
東京都千代田区大手町二丁目３番１号
(72)発明者　巌寺　俊哲
東京都新宿区西新宿三丁目19番２号　日本電信電話株式会社内
(74)代理人　弁理士　伊東　忠彦

(54)【発明の名称】　文書情報抽出方法及び装置及び文書情報抽出プログラムを格納した記憶媒体

(57)【要約】
【課題】　予め、パターンやヒューリスティックスなどを用意することなく、新たな情報の抽出も可能な文書情報抽出方法及び装置及び文書情報抽出プログラムを格納した記憶媒体を提供する。
【解決手段】　本発明は、対象文書集合を標準文書集合に対して特徴付ける第１の特徴情報を、該対象文書集合中から算出し、対象文書集合中の各個別文書を他の個別文書に対して特徴付ける第２の特徴情報を該対象文書集合中の各個別文書から算出し、第１の特徴情報と、前記第２の特徴情報に基づいて、前記対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出し、抽出された前記情報を各個別文書を特徴づける情報として出力する。

本発明の原理を説明するための図



1

【特許請求の範囲】
【請求項１】　文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出方法において、
前記対象文書集合を標準文書集合に対して特徴付ける第１の特徴情報を、該対象文書集合中から算出し、
前記対象文書集合中の各個別文書について、各個別文書を他の個別文書に対して特徴付ける第２の特徴情報を該対象文書集合中の各個別文書から算出し、
前記第１の特徴情報と、前記第２の特徴情報に基づいて、前記対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出し、
抽出された前記情報を各個別文書を特徴づける情報として出力することを特徴とする文書情報抽出方法。
【請求項２】　文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出装置であって、
前記対象文書集合を標準文書集合に対して特徴付ける第１の特徴情報を、該対象文書集合中から算出する第１の特徴情報算出手段と、
前記対象文書集合中の各個別文書について、他の個別文書を特徴付ける第２の特徴情報を該対象文書集合中の各個別文書から算出する第２の特徴情報算出手段と、
前記第１の特徴情報算出手段で算出された前記第１の特徴情報と、前記第２の特徴情報算出手段で算出された前記第２の特徴情報に基づいて、前記対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出する個別文書特徴抽出手段と、
抽出された前記情報を各個別文書を特徴づける情報として出力する特徴情報出力手段とを有することを特徴とする文書情報抽出装置。
【請求項３】　前記入力記憶手段から標準文書集合を受け取る標準文書集合更新手段と、
前記標準文書集合更新手段に与えられた前記標準文書集合中の各文書を解析し、該文書を構成する単語と該単語の標準文書集合中での出現頻度を算出する標準文書集合解析手段と、
前記標準文書集合中の単語と該単語の出現頻度を対応付けて記憶する標準文書集合解析結果記憶手段と、
前記入力記憶手段から複数の文書で構成される対象文書集合を受け取る対象文書集合入力手段と、
前記対象文書集合中の各文書を解析し、該文書の各個別文書を構成する単語と該単語の該文書中での出現頻度を算出する対象文書集合解析手段と、
前記各対象文書集合中の単語と該単語の出現頻度を各文書を対応付けて記憶する対象文書集合解析結果記憶手段

2

と、
前記各対象文書集合中の各個別文書中の単語と該単語の出現頻度を該単語が出現した文書と対応付けて記憶する個別文書解析結果記憶手段と、
前記対象文書集合全体としての特徴情報を、前記対象文書集合解析結果記憶手段に記憶されている情報を用いて算出する対象文書集合全体特徴算出手段と、
前記対象文書集合全体特徴算出手段によって算出された、前記対象文書集合全体としての特徴情報を記憶する対象文書集合全体特徴記憶手段と、
前記対象文書集合中の各個別文書の特徴情報を、前記個別文書解析結果記憶手段に記憶されている情報を用いて算出する個別文書特徴算出手段と、
前記個別文書特徴算出手段によって算出された、前記対象文書集合中の各個別文書に対応する特徴情報を記憶する個別文書特徴記憶手段と、
前記個別文書解析結果記憶手段または、前記対象文書集合解析結果記憶手段に記憶されているデータを一時的に記憶する目的情報一時記憶手段と、
前記対象文書集合解析結果記憶手段または、前記標準文書集合解析結果記憶手段に記憶されているデータを一時的に記憶する基準情報一時記憶手段と、
前記目的情報一時記憶手段に記憶されているデータと前記基準情報一時記憶手段に記憶されているデータとを比較し、該目的情報一時記憶手段に記憶されているデータの特徴スコアを算出する目的情報特徴スコア算出手段と、
前記対象文書集合全体特徴記憶手段に記憶されているデータと前記個別文書特徴記憶手段に記憶されているデータを用いて、各個別文書の特徴情報を算出する特徴情報算出手段と、
前記特徴情報算出手段において算出された各個別文書の特徴情報を記憶する特徴情報記憶手段と、
前記特徴情報記憶手段に記憶されている各個別文書の特徴情報を用いて、前記対象文書集合中の各個別文書から特徴表現を抽出する特徴表現抽出手段と、
前記特徴表現抽出手段により前記各個別文書から抽出された特徴表現を記憶する特徴表現記憶手段と、
前記特徴表現記憶手段に記憶されている特徴表現を転送媒体に与える特徴表現出力手段とを有する請求項２記載の文書情報抽出装置。
【請求項４】　前記特徴情報算出手段は、
前記対象文書集合全体特徴と前記個別文書特徴の特徴スコアを掛けた数値を用いる請求項３記載の文書情報抽出装置。
【請求項５】　前記特徴情報算出手段は、
前記特徴情報を算出する際にχ２乗検定を用いる請求項４記載の文書情報抽出装置。
【請求項６】　前記特徴表現抽出手段は、
前記対象文書集合中の各文書中の全単語に、単語の特徴

を数値化した特徴情報スコアを付与する特徴スコア付与手段と、
前記各文書中に含まれる予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた文書中の部分構造を構成する単語列である各文毎に、該文を構成する単語に付与されている前記特徴情報スコアの平均を求める平均算出手段と、
前記平均算出手段により求められた前記平均の値が最大の文書内部分表現を前記文書の特徴表現として抽出する特徴表現決定手段とを含む請求項３記載の文書情報抽出装置。
【請求項７】　文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出プログラムを格納した記憶媒体であって、
前記対象文書集合を標準文書集合に対して特徴付ける第１の特徴情報を、該対象文書集合中から算出する第１の特徴情報算出プロセスと、
前記対象文書集合中の各個別文書を、他の個別文書に対して特徴付ける第２の特徴情報を該対象文書集合中の各個別文書から算出する第２の特徴情報算出プロセスと、
前記第１の特徴情報算出プロセスで算出された前記第１の特徴情報と、前記第２の特徴情報算出プロセスで算出された前記第２の特徴情報に基づいて、前記対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出する個別文書特徴抽出プロセスと、
抽出された前記情報を各個別文書を特徴づける情報として出力する特徴情報出力プロセスとを有することを特徴とする文書情報抽出プログラムを格納した記憶媒体。
【請求項８】　前記入力記憶手段から標準文書集合を受け取る標準文書集合更新プロセスと、
前記標準文書集合更新プロセスに与えられた前記標準文書集合中の各文書を解析し、該文書を構成する単語と該単語の標準文書集合中での出現頻度を算出する標準文書集合解析プロセスと、
前記入力記憶手段から複数の文書で構成される対象文書集合を受け取る対象文書集合入力プロセスと、
前記対象文書集合中の各文書を解析し、該文書の各個別文書を構成する単語と該単語の該文書中での出現頻度を算出する対象文書集合解析プロセスと、
前記対象文書集合全体としての特徴情報を、前記各対象文書中の単語と該単語の出現頻度が記憶されている対象文書集合解析結果記憶手段の情報及び標準文書集合解析結果記憶手段の情報を用いて算出する対象文書集合全体特徴算出プロセスと、
前記対象文書集合中の各個別文書の特徴情報を、前記各対象文書集合中の各個別文書中の単語と該単語の出現頻度を該単語が出現した文書と対応付けて記憶されている個別文書解析結果記憶手段の情報及び対象文書集合解析結果記憶手段の情報を用いて算出する個別文書特徴算出プロセスと、
前記個別文書の解析結果または、前記対象文書集合解析結果を一時的に記憶している目的情報一時記憶手段に記憶されているデータと、前記対象文書集合解析プロセスまたは、前記標準文書集合解析プロセスの結果を一時的に記憶している基準情報一時記憶手段のデータとを比較し、該目的情報一時記憶手段に記憶されているデータの特徴スコアを算出する目的情報特徴スコア算出プロセスと、
前記対象文書集合全体特徴算出プロセスによって算出された、前記対象文書集合全体としての特徴情報を記憶している対象文書集合全体特徴記憶手段のデータと、前記個別文書特徴算出プロセスによって算出された、前記対象文書集合中の各個別文書の特徴情報を記憶している個別文書特徴記憶手段のデータを用いて、各個別文書の特徴情報を算出する特徴情報算出プロセスと、
前記特徴情報算出プロセスにおいて算出された各個別文書の特徴情報が記憶されている特徴情報記憶手段の各個別文書の特徴情報を用いて、前記対象文書集合中の各個別文書から特徴表現を抽出する特徴表現抽出プロセスと、
前記特徴表現抽出プロセスにより前記各個別文書から抽出された特徴表現が記憶されている特徴表現記憶手段の特徴表現を転送媒体に与える特徴表現出力プロセスとを有する請求項７記載の文書情報抽出プログラムを格納した記憶媒体。
【請求項９】　前記特徴情報算出プロセスは、
前記対象文書集合全体特徴と前記個別文書特徴の特徴スコアを掛けた数値を用いる請求項８記載の文書情報抽出プログラムを格納した記憶媒体。
【請求項１０】　前記特徴情報算出プロセスは、
前記特徴情報を算出する際にχ２乗検定を用いる請求項９記載の文書情報抽出プログラムを格納した記憶媒体。
【請求項１１】　前記特徴表現抽出プロセスは、
前記対象文書集合中の各文書中の全単語に、単語の特徴を数値化した特徴情報スコアを付与する特徴スコア付与プロセスと、
前記各文書中に含まれる予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた文書中の部分構造を構成する単語列である各文毎に、該文を構成する単語に付与されている前記特徴情報スコアの平均を求める平均算出プロセスと、
前記平均算出プロセスにより求められた前記平均の値が最大の文書内部分表現を前記文書の特徴表現として抽出する特徴表現決定プロセスとを含む請求項８記載の文書情報抽出プログラムを格納した記憶媒体。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、文書情報抽出方法

及び装置及び文書情報抽出プログラムを格納した記憶媒体に係り、特に、文書情報処理に用いられる文書情報抽出処理において、特定の話題についての文書集合中の各文書より、当該話題に関連し、かつ、文書集合中の他文書との相違をより明確に示す情報を抽出する文書情報抽出方法及び装置及び文書情報抽出プログラムを格納した記憶媒体に関する。

【０００２】

【従来の技術】近年、インターネットが急速に普及している。さらに、データ記憶装置は、大容量化、低価格化している。これに伴って大量で多様な情報が、ネットワークを介して容易に利用可能になっている。また、WWWの普及と共に多くのユーザが相互に情報を生成し、利用している。しかし、情報洪水と言われるように利用できる情報量が飛躍的に増加するに従って、これらの情報の中から有益な情報を見つけ出して取捨選択することが困難になってきている。

【０００３】このような大量の情報を全て閲覧し、有益な情報を探索し、選別することは困難である。従って、適切な情報を効率的に利用するためには、大量の情報から特徴的な情報を抽出し、必要十分な情報を選択的に利用可能にする必要がある。現在、情報を選択的に利用する手段として情報検索技術が用いられている。しかし、ネットワークを介して利用できる情報は、その分量が膨大であり、その内容も多岐に渡っている。このため、一度の検索結果として多くの類似の情報を含む文書が選択されてしまう。

【０００４】さらに、検索結果から適切な情報を含む文書を選択することを支援するために、多くの場合、各文書を特徴付ける部分情報を文書から抽出し、各検索結果文書に付与し、利用者に提示している。膨大な数の文書の各文書から適量・適質な特徴的な部分情報の、人手による抽出は、困難である。また、人手により作業を行った場合、抽出される情報は、作業者のもっている主観や知識に影響されるため、複数の作業者によって抽出が行われると抽出された情報の品質を均質に保つことができない。そこで、情報を選択的に利用するためには、文書内容から適切な部分情報を自動的に決定し、抽出する情報抽出技術が必要となる。

【０００５】従来の情報抽出技術の代表的なものとして次の２つが挙げられる。第１には、パターンマッチングによる情報抽出であり、第２には、ヒューリスティックスを用いた情報抽出技術がある。パターンマッチングによる情報抽出技術は、ある単語列をパターンとして予め保持しておき、パターンマッチング処理によって情報を抽出する技術である。これは、特定の情報が、限られたパターンによって表現されることが多いという考え方に基づいている。

【０００６】例えば、予め、「＜メーカ＞は、＜製品＞の販売を開始した」のようなパターンを用意しておくことにより、文書中からこのパターンにマッチする「○×コンピュータは新型コンピュータの販売を開始した。」という文を抽出する。また、ヒューリスティックスを用いた情報抽出技術は、文の文書中での位置情報、タイトルや見出しに出現する単語、手がかり語句の有無を組み合わせて文の重要度を判定し、重要と判定された文を抽出する技術である。

【０００７】

【発明が解決しようとする課題】しかしながら、上記従来のパターンマッチングによる情報抽出技術は、必要とするパターンを予め用意しておくことが必要である。このため、パターンマッチングしない新たな情報の抽出を行うことができないという問題がある。また、同一の文書からは、必ず同一の部分が抽出される。

【０００８】また、上記従来のヒューリスティックスを用いた情報抽出技術は、前述のパターンマッチングによる情報抽出技術と同様に、予めヒューリスティックスを用意しておくことが必要である。このため、新たな情報の抽出には、不向きである。また、同一の文書からは、必ず同一の部分が抽出される。さらに、ある文書タイプで有効なヒューリスティックスが別の文書タイプで有効であるとは限らない。例えば、新聞記事などでは、位置情報が有効である。また、学術論文では、手がかり語句によるヒューリスティックスが有効である。インターネット上には、様々なタイプの文書が混在しており、文書タイプを自動的に判定することが必要となる。しかし、現状では、文書タイプを判別する有効な技術がない、という問題がある。

【０００９】また、タイトルや見出しが存在しない文書や、手がかり語句が殆ど出現しない文書も多いため、ヒューリスティックスが有効に働かないことが多いという問題がある。本発明は、上記の点に鑑みなされたもので、予め、パターンやヒューリスティックスなどを用意することなく、新たな情報の抽出も可能な文書情報抽出方法及び装置及び文書情報抽出プログラムを格納した記憶媒体を提供することを目的とする。

【００１０】

【課題を解決するための手段】図１は、本発明の原理を説明するための図である。本発明（請求項１）は、文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出方法において、対象文書集合を標準文書集合に対して特徴付ける第１の特徴情報を、該対象文書集合中から算出し（ステップ１）、対象文書集合中の各個別文書について、各個別文書を他の個別文書に対して特徴付ける第２の特徴情報を該対象文書集合中の各個別文書から算出し（ステップ２）、第１の特徴情報と、第２の特徴情報に基づいて、対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書か

ら抽出し（ステップ３）、抽出された情報を各個別文書を特徴づける情報として出力する（ステップ４）。

【００１１】図２は、本発明の原理構成図である。本発明（請求項２）は、文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出装置であって、対象文書集合を標準文書集合に対して特徴付ける第１の特徴情報を、該対象文書集合中から算出する第１の特徴情報算出手段１と、対象文書集合中の各個別文書を、他の個別文書に対して特徴付ける第２の特徴情報を該対象文書集合中の各個別文書から算出する第２の特徴情報算出手段２と、第１の特徴情報算出手段１で算出された第１の特徴情報と、第２の特徴情報算出手段２で算出された第２の特徴情報に基づいて、対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出する個別文書特徴抽出手段３と、抽出された情報を各個別文書を特徴づける情報として出力する特徴情報出力手段４とを有する。

【００１２】本発明（請求項３）は、入力記憶手段から標準文書集合を受け取る標準文書集合更新手段と、標準文書集合更新手段に与えられた標準文書集合中の各文書を解析し、該文書を構成する単語と該単語の標準文書集合中での出現頻度を算出する標準文書集合解析手段と、標準文書集合中の単語と該単語の出現頻度を対応付けて記憶する標準文書集合解析結果記憶手段と、入力記憶手段から複数の文書で構成される対象文書集合を受け取る対象文書集合入力手段と、対象文書集合中の各文書を解析し、該文書の各個別文書を構成する単語と該単語の該文書中での出現頻度を算出する対象文書集合解析手段と、各対象文書集合中の単語と該単語の出現頻度を対応付けて記憶する対象文書集合解析結果記憶手段と、各対象文書集合中の各個別文書中の単語と該単語の出現頻度を該単語が出現した文書と対応付けて記憶する個別文書解析結果記憶手段と、対象文書集合全体としての特徴情報を、対象文書集合解析結果記憶手段及び標準文書集合解析結果記憶手段に記憶されている情報を用いて算出する対象文書集合全体特徴算出手段と、対象文書集合全体特徴算出手段によって算出された、対象文書集合全体としての特徴情報を記憶する対象文書集合全体特徴記憶手段と、対象文書集合中の各個別文書の特徴情報を、個別文書解析結果記憶手段及び対象文書集合解析結果記憶手段に記憶されている情報を用いて算出する個別文書特徴算出手段と、個別文書特徴算出手段によって算出された、対象文書集合中の各個別文書に対応する特徴情報を記憶する個別文書特徴記憶手段と、個別文書解析結果記憶手段または、対象文書集合解析結果記憶手段に記憶されているデータを一時的に記憶する目的情報一時記憶手段と、対象文書集合解析結果記憶手段または、標準文書集合解析結果記憶手段に記憶されているデータを一時的

に記憶する基準情報一時記憶手段と、目的情報一時記憶手段に記憶されているデータと基準情報一時記憶手段に記憶されているデータとを比較し、該目的情報一時記憶手段に記憶されているデータの特徴スコアを算出する目的情報特徴スコア算出手段と、対象文書集合全体特徴記憶手段に記憶されているデータと個別文書特徴記憶手段に記憶されているデータを用いて、各個別文書の特徴情報を算出する特徴情報算出手段と、特徴情報算出手段において算出された各個別文書の特徴情報を記憶する特徴情報記憶手段と、特徴情報記憶手段に記憶されている各個別文書の特徴情報を用いて、対象文書集合中の各個別文書から特徴表現を抽出する特徴表現抽出手段と、特徴表現抽出手段により各個別文書から抽出された特徴表現を記憶する特徴表現記憶手段と、特徴表現記憶手段に記憶されている特徴表現を転送媒体に与える特徴表現出力手段とを有する。

【００１３】本発明（請求項４）は、特徴情報算出手段において、対象文書集合全体特徴と個別文書特徴の特徴スコアを掛けた数値を用いる。本発明（請求項５）は、特徴情報算出手段において、特徴情報を算出する際にχ２乗検定を用いる。本発明（請求項６）は、特徴表現抽出手段において、対象文書集合中の各文書中の全単語に、単語の特徴を数値化した特徴情報スコアを付与する特徴スコア付与手段と、各文書中に含まれる予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた文書中の部分構造を構成する単語列である各文毎に、該文を構成する単語に付与されている特徴情報スコアの平均を求める平均算出手段と、平均算出手段により求められた平均の値が最大の文書内部分表現を文書の特徴表現として抽出する特徴表現決定手段とを含む。

【００１４】本発明（請求項７）は、文書データを記憶した入力記憶手段から読み出される複数の文書から構成される対象文書集合中の各個別文書を特徴付ける情報を抽出する文書情報抽出プログラムを格納した記憶媒体であって、対象文書集合を標準文書集合に対して特徴付ける第１の特徴情報を、該対象文書集合中から算出する第１の特徴情報算出プロセスと、対象文書集合中の各個別文書を、他の個別文書に対して特徴付ける第２の特徴情報を該対象文書集合中の各個別文書から算出する第２の特徴情報算出プロセスと、第１の特徴情報算出プロセスで算出された第１の特徴情報と、第２の特徴情報算出プロセスで算出された第２の特徴情報に基づいて、対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を該各個別文書から抽出する個別文書特徴抽出プロセスと、抽出された情報を各個別文書を特徴づける情報として出力する特徴情報出力プロセスとを有する。

【００１５】本発明（請求項８）は、入力記憶手段から標準文書集合を受け取る標準文書集合更新プロセスと、

標準文書集合更新プロセスに与えられた標準文書集合中の各文書を解析し、該文書を構成する単語と該単語の標準文書集合中での出現頻度を算出する標準文書集合解析プロセスと、入力記憶手段から複数の文書で構成される対象文書集合を受け取る対象文書集合入力プロセスと、対象文書集合中の各文書を解析し、該文書の各個別文書を構成する単語と該単語の出現頻度を算出する対象文書集合解析プロセスと、対象文書集合全体としての特徴情報を、各対象文書中の単語と該単語の出現頻度が記憶されている対象文書集合解析結果記憶手段の情報及び標準文書集合解析結果記憶手段の情報を用いて算出する対象文書集合全体特徴算出プロセスと、対象文書集合中の各個別文書の特徴情報を、各対象文書集合中の各個別文書中の単語と該単語の出現頻度を該単語が出現した文書と対応付けて記憶されている個別文書解析結果記憶手段の情報及び対象文書解析結果記憶手段の情報を用いて算出する個別文書特徴算出プロセスと、個別文書の解析結果または、対象文書集合解析結果を一時的に記憶している目的情報一時記憶手段に記憶されているデータと、対象文書集合解析プロセスまたは、標準文書集合解析プロセスの結果を一時的に記憶している基準情報一時記憶手段のデータとを比較し、該目的情報一時記憶手段に記憶されているデータの特徴スコアを算出する目的情報特徴スコア算出プロセスと、対象文書集合全体特徴算出プロセスによって算出された、対象文書集合全体としての特徴情報を記憶している対象文書集合全体特徴記憶手段のデータと、個別文書特徴算出プロセスによって算出された、対象文書集合中の各個別文書の特徴情報を記憶している個別文書特徴記憶手段のデータを用いて、各個別文書の特徴情報を算出する特徴情報算出プロセスと、特徴情報算出プロセスにおいて算出された各個別文書の特徴情報が記憶されている特徴情報記憶手段の各個別文書の特徴情報を用いて、対象文書集合中の各個別文書から特徴表現を抽出する特徴表現抽出プロセスと、特徴表現抽出プロセスにより各個別文書から抽出された特徴表現が記憶されている特徴表現記憶手段の特徴表現を転送媒体に与える特徴表現出力プロセスとを有する。

【００１６】本発明（請求項９）は、特徴情報算出プロセスにおいて、対象文書集合全体特徴と個別文書特徴の特徴スコアを掛けた数値を用いる。本発明（請求項１０）は、特徴情報算出プロセスにおいて、特徴情報を算出する際にχ２乗検定を用いる。本発明（請求項１１）は、特徴表現抽出プロセスにおいて、対象文書集合中の各文書中の全単語に、単語の特徴を数値化した特徴情報スコアを付与する特徴スコア付与プロセスと、各文書中に含まれる予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた文書中の部分構造を構成する単語列である各文毎に、該文を構成する単語に付与されている特徴情報スコアの平均を求める平均算出プロセスと、平均算出プロセスにより求められた平均の値が最大の文書内部分表現を文書の特徴表現として抽出する特徴表現決定プロセスとを含む。

【００１７】上述のように、本発明は、対象文書集合を標準文書集合に対して特徴付ける情報を対象文書集合中から算出し、次に、対象文書集合中の各個別文書について、他の個別文書に対してその文書を特徴付ける情報を個別文書から算出し、これらの処理により得られた特徴情報に基づいて、対象文書集合をより特徴付ける情報であり、かつ、各個別文書を他の個別文書に対して特徴付ける情報を各個別文書から抽出する。抽出された情報を各個別文書を特徴付ける情報として出力する。

【００１８】これにより、例えば、「桜の花見」について検索を実行した結果、得られる文書集合を「対象文書集合」とし、この文書検索装置が検索対象とする文書集合全体を「標準文書集合」とした場合を想定すると、この場合、対象文書集合を標準文書集合に対して特徴付ける情報は、「桜の花見」に関する情報、例えば、桜の見頃や名所に関する情報になる。

【００１９】また、各個別文書を他の個別文書に対して特徴付ける情報は、他の文書と差異がある情報、例えば、「４月の上旬」のような時間的な情報や、「上野公園」のような場所的情報になる。これらの情報を組み合わせて用いることにより、各個別文書からその文書を特徴付ける表現、例えば、「…見頃は、４月上旬…」、「…名所：上野公園…」のような時間的、場所的記述を表している部分を抽出することが可能となる。

【００２０】このように、本発明では、大量の文書情報から各文書を特徴付ける情報を適時に適質適量抽出可能となる。また、予め情報抽出に使用する知識等を用意する必要がなく、様々な文書から情報抽出に適用可能である。

【００２１】

【発明の実施の形態】図３は、本発明の文書情報抽出装置の構成を示す。同図に示す構成は、例えば、文書検索システムの一部を構成し、検索された文書集合中の各文書から各文書間の相違を明確に示し、各文書を特徴付ける適切な情報を抽出し、提示することにより、使用者が必要とする文書を選択することを支援する装置である。以下、文書検索システム本体から出力された文書集合中の各文書から情報抽出する場合を想定して説明する。ここで、当該文書情報抽出装置は、入力として文書集合を受信し、予め、提供されている初期情報を用いて入力された文書集合を処理し、各文書毎に情報を抽出し、出力するものである。

【００２２】同図に示す構成は、監視制御部１０、入力記憶装置２０、転送媒体３０、各種制御処理を実行する処理部１０１～１１０及び各種データを記憶する記憶部２０１～２０９から構成される。処理部は、標準文書集合更新部１０１、標準文書集合解析部１０２、対象文書

集合入力部１０３、対象文書集合解析部１０４、対象文書集合全体特徴算出部１０５、個別文書特徴算出部１０６、特徴情報算出部１０７、特徴表現抽出部１０８、特徴表現出力部１０９、目的情報特徴スコア算出部１１０から構成される。

【００２３】記憶部は、標準文書集合解析結果記憶部２０１、対象文書集合解析結果記憶部２０２、対象文書集合全体特徴記憶部２０３、個別文書解析結果記憶部２０４、個別文書特徴記憶部２０５、特徴情報記憶部２０６、特徴表現記憶部２０７、基準情報一時記憶部２０８、及び目的情報一時記憶部２０９から構成される。各処理部１０１～１１０を総合的に監視制御する監視制御部１０に、処理部１０１～１０９と、記憶部２０１～２０７が接続される。また、対象文書集合全体特徴算出部１０５と個別文書特徴算出部１０６には、目的情報特徴スコア算出部１１０、基準情報一時記憶部２０８及び目的情報一時記憶部２０９が接続される。さらに、基準情報一時記憶部２０８、目的情報一時記憶部２０９は、目的特徴スコア算出部１１０にも接続される。

【００２４】ここで、各処理部は、例えば、ディジタル電子計算機で構成され、それぞれＣＰＵと、動作プログラムとそれを実行するためのデータを記録するＲＯＭと、ワーキングメモリとして用いられるＲＡＭとを備える。なお、全処理部を１つのディジタル電子計算機で構成してもよい。さらに、各記憶部２０１～２０９は、例えば、ハードディスクメモリなどのメモリに記憶される。

【００２５】また、入力記憶部２０には、本装置に与えられる標準文書集合、対象文書集合が一定の順序で記憶されている。入力記憶部２０は、半導体メモリ装置、あるいは、ハードディスクやフロッピーディスクによって実現することができる。転送媒体３０には、本装置の処理結果が与えられる通信チャネルまたは記録媒体である。

【００２６】以下の説明において、「標準文書集合」とは、特徴情報抽出の対象となり得る文書全体集合、または、文書全体集合を母集合とし、その標本集合となる文書集合を指す。また、「対象文書集合」とは、特徴情報抽出の対象の文書集合を指す。さらに、「個別文書」とは、「対象文書集合」中の各文書を指す。

【００２７】「特徴表現」とは、各「個別文書」から抽出される表現であり、「対象文書集合」中の他の文書に対して、当該「個別文書」を特徴付ける表現である。標準文書集合更新部１０１は、入力記憶装置２０から標準文書集合を受け取る。標準文書集合解析部１０２は、標準文書集合更新部１０１に与えられた標準文書集合中の各文書を解析し、その文を構成する単語とその単語の標準文書集合中での出現頻度を算出する。

【００２８】標準文書集合解析結果記憶部２０１は、標準文書集合中の単語とその単語の出現頻度を対応付けて

記憶する。対象文書集合入力部１０３は、入力記憶装置２０から複数の文書で構成される対象文書集合を受け取る。対象文書集合解析部１０４は、対象文書集合中の各文書を解析し、その各個別文書を構成する単語と当該単語の出現頻度を算出する。

【００２９】対象文書集合解析結果記憶部２０２は、対象文書集合解析部１０４で求められた各対象文書中の単語と単語の出現頻度を各文書と対応付けて記憶する。個別文書解析結果記憶部２０４は、対象文書集合解析部１０４で求められた対象文書集合中の各個別文書中の単語とその単語の出現頻度をその単語が出現した文書と対応付けて記憶する。

【００３０】対象文書集合全体特徴算出部１０５は、対象文書集合全体としての特徴情報を、上記対象文書集合解析結果記憶部２０２に記憶されている情報及び上記標準文書集合解析結果記憶部２０１に記憶されている情報を用いて算出する。対象文書集合全体特徴記憶部２０３は、対象文書集合全体特徴算出部１０５によって算出された、対象文書集合全体としての特徴情報を記憶する。

【００３１】個別文書特徴算出部１０６は、対象文書集合中の各個別文書の特徴情報を個別文書解析結果記憶部２０４に記憶されている情報及び対象文書集合解析結果記憶部２０２に記憶されている情報を用いて算出する。個別文書特徴記憶部２０５は、個別文書特徴算出部１０６によって算出された対象文書集合中の各個別文書の特徴情報を記憶する。

【００３２】目的情報一時記憶部２０９は、個別文書解析結果または、対象文書集合解析結果を一時的に記憶する。基準情報一時記憶部２０８は、対象文書集合解析結果、または、標準文書集合解析結果を一時的に記憶する。目的情報特徴スコア算出部１１０は、目的情報一時記憶部２０９に記憶されているデータと上記基準情報一時記憶部２０８に記憶されているデータを比較し、上記の目的情報一時記憶部２０９に記憶されているデータの特徴スコアを算出する。

【００３３】特徴情報算出部１０７は、対象文書集合全体特徴記憶部２０３に記憶されているデータと個別文書特徴記憶部２０５に記憶されているデータを用いて各個別文書の特徴情報を算出する。特徴情報記憶部２０６は、特徴情報算出部１０７において算出された各個別文書の特徴情報を記憶する。

【００３４】特徴表現抽出部１０８は、特徴情報記憶部２０６に記憶されている各個別文書の特徴情報を用いて、対象文書集合中の各個別文書から特徴表現を抽出する。特徴表現記憶部２０７は、特徴表現抽出部１０８により各個別文書から抽出された特徴表現を記憶する。特徴表現出力部１０９は、特徴表現記憶部２０７に記憶されている特徴表現を転送媒体３０に与える。

【００３５】

【実施例】以下、図面と共に本発明の実施例を説明す

る。まず、監視制御部１０及び対象文書集合全体特徴算出部１０５、個別文書特徴算出部１０６に接続される記憶部２０１～２０９について説明する。標準文書集合解析結果記憶部２０１は、標準文書集合解析部１０２の処理結果である、標準文書集合の解析結果を記憶・保持する。標準文書集合の解析結果とは、標準文書集合中の全文書に記述されている文章を形態素解析し、各単語の表現及び各単語出現頻度を対応付けたものである。解析結果は、単語表現、出現頻度の２つのカラムからなるテーブルとして表現・記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。

【００３６】対象文書集合解析結果記憶部２０２は、対象文書集合解析部１０４の処理結果の１つとして得られる対象文書集合の解析結果を記憶・保持する。対象文書集合とは、本装置が接続される情報検索装置において検索作業の実行を結果として得られる文書集合である。対象文書集合の解析結果とは、対象文書集合中の全文書に記述されている文章を形態素解析し、各単語の表現と対象文書集合中での各単語の出現頻度を対応付けたものである。解析結果は、標準文書集合の解析結果と同様の形式であり、単語表現、出現頻度の２つのカラムからなるテーブルとして、表現、記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。

【００３７】個別文書解析結果記憶部２０４は、対象文書集合解析部１０４の処理結果の１つとして得られる個別文書の解析結果を記憶・保持する。ここで、個別文書とは、本装置が接続される情報検索装置において、検索作業の実行の結果として得られる文書集合、即ち、前述した対象文書集合に含まれる各文書である。個別文書の解析結果とは、対象文書集合中の各文書毎に記述されている文章を形態素解析し、各単語の表現、及びその文書中での各単語の出現頻度を対応付け、文書毎に記録したものである。解析結果は、標準文書集合の解析結果と同様の形式であり、単語表現、出現頻度の２つのカラムからなるテーブルとして表現、記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。また、対象文書集合中の各文書毎に１個のテーブルが構成される。

【００３８】対象文書集合全体特徴記憶部２０３は、対象文書集合全体特徴算出部１０５の処理結果として得られる、標準文書集合に対する対象文書集合全体として特徴を点数化した情報を記憶・保持する。ここで、対象文書集合全体特徴とは、標準文書集合中の単語の出現頻度分布と対象文書集合中の単語の出現頻度分布を比較し、

その分布の相違の大小を各単語毎に数値化したものであり、標準文書集合中の出現頻度分布と対象文書集合中の出現頻度分布の相違が大きい単語ほど大きな数値をとる。対象文書集合中で特徴的な単語、即ち、標準文書集合中の出現頻度分布と対象文書集合中の出現分布の相違が大きい単語ほど大きな数値をとる。対象文書集合全体特徴は、各単語表現とその単語の出現分布の特徴を数値化し、表現した特徴スコアの２つのカラムからなるテーブルとして表現される。各行は、各単語の表現と特徴スコアの対応を表す。対象文書集合全体に対して１つのテーブルが対応する。

【００３９】個別文書特徴記憶部２０５は、個別文書特徴算出部１０６の処理結果として得られる。対象文書集合全体に対する対象文書集合中の各個別文書の特徴を点数化し、記憶・保持する。ここで、個別文書特徴とは、対象文書集合中の単語の出現頻度分布と対象文書集合中の各個別文書中の単語の出現頻度分布を比較し、その分布の相違の大小を各単語毎に数値化したものであり、対象文書集合中の出現分布と個別文書中の出現分布の相違が大きい単語ほど、大きな数値をとる。即ち、個別文書（対象文書集合中の各文書）に特徴的な単語ほどより大きな数値を持つ。個別文書特徴は、各単語表現とその単語の出現分布の特徴を数値化し表現した特徴スコアの２つのカラムからなるテーブルとして表現される。各行は、各単語の表現と特徴スコアの対応を表す。また、個別文書毎に１つのテーブルが対応する。

【００４０】特徴情報記憶部２０６は、特徴情報算出部１０７の処理結果として得られる、各文書の特徴情報をその情報の算出元である文書と対応付けて記憶・保持する。ここで、特徴情報とは、各単語毎に対象文書集合中での特徴スコアと個別文書中での特徴スコアをかけた数値であり、対象文書集合中に特徴的単語であり、かつ個別文書中でも特徴的な単語ほど大きな数値を持つ。特徴情報は、各単語表現とその単語の特徴を数値化した特徴スコアの２つのカラムからなるテーブルとして表現される。各行は、各単語表現と特徴スコアの対応を表す。また、個別文書毎に１つのテーブルが対応する。

【００４１】特徴表現記憶部２０７は、特徴表現抽出部１０８の処理結果として得られる、対象文書集合中の各文書毎の特徴表現をその情報の抽出元である文書と対応付けて記憶・保持する。各文書の特徴表現とは、各文書に含まれる、予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた部分構造を構成する単語列であり、その連続する単語列、文、部分構造を構成する単語列を構成する単語全体の特徴スコアの平均が最大の部分である。

【００４２】基準情報一時記憶部２０８は、目的情報特徴スコア算出部１１０における特徴スコアの算出に用いる基準情報を一時的に記憶・保持する。基準情報は、単語表現、出現頻度の２つのカラムからなるテーブルとし

て表現、記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。この記憶部２０８には、一度に上記の形式のテーブルが１個、記憶・保持される。

【００４３】目的情報一時記憶部２０９は、目的情報特徴スコア算出部１１０における特徴スコアの算出において、特徴スコアの算出の対象となる目的情報を一時的に記憶・保持する。目的情報は、単語表現、出現頻度の２つのカラムからなるテーブルとして表現、記憶・保持される。このテーブルにおいて、各行は、各単語表現とその単語の出現頻度の対応関係を表す。このテーブルは、各単語表現をキーとして対応する行を検索できる構造をとる。この記憶部２０９には、一度に上記形式のテーブルが一個、記憶・保持される。

【００４４】次に、図３に示す各処理部について説明する。監視制御部１０は、処理部１０１～１０９を制御し、データフローを統制するモジュールである。図４は、本発明の一実施例の文書情報抽出処理のフローチャートである。以下、同図のフローチャートに沿って、各処理部の動作を説明する。

【００４５】ステップ１０１）　監視制御部１０において、標準文書集合が更新されているか否かが判断される。更新された場合には、ステップ１０２に移行し、更新されていない場合は、ステップ１０５に移行する。

ステップ１０２）　監視制御部１０は、更新された標準文書集合を入力記憶装置２０から標準文書集合更新部１０１へ転送する。

【００４６】ステップ１０３）　この時点で、標準文書集合更新部１０１は、転送された標準文書集合に対して、標準文書集合更新処理を実行し、処理結果である標準文書集合更新結果を監視制御部１０に出力する。監視制御部１０は、標準文書集合更新部１０１から出力されたすべての標準文書集合更新結果を標準文書集合解析部１０２へ転送する。

【００４７】ステップ１０４）　標準文書集合解析部１０２は、標準文書集合解析処理を実行し、処理結果を監視制御部１０へ出力する。これにより、監視制御部１０は、標準文書集合解析部１０２から出力されたすべての標準文書集合解析結果を標準文書集合解析結果記憶部２０１に転送し、その内容を更新し、新たに転送された値を記憶・保持する。

【００４８】ステップ１０５）　監視制御部１０によって対象文書集合が入力されたか否かが判断される。入力された場合は、ステップ１０６へ移行し、入力されていない場合には、ステップ１０５の処理を繰り返す。

ステップ１０６）　監視制御部１０は、入力された対象文書集合を入力記憶装置２０から対象文書集合入力部１０３へ転送する。対象文書集合入力部１０３は、入力された対象文書集合に対して対象文書集合入力処理を実行し、処理結果を監視制御部１０に出力する。監視制御部１０は、対象文書集合入力部１０３から出力される対象文書集合入力処理結果を対象文書集合解析部１０４へ転送する。

【００４９】ステップ１０７）　対象文書集合解析部１０４は、対象文書集合解析処理を実行し、解析結果を監視制御部１０へ出力する。

ステップ１０８）　監視制御部１０は、対象文書集合解析部１０４から出力された対象文書集合解析結果を対象文書集合解析結果記憶部２０２に転送すると共に、個別文書解析結果を個別文書解析結果記憶部２０３に転送し、各記憶部の内容を更新し、新たに転送された値を記憶・保持する。さらに、監視制御部１０は、標準文書集合解析結果記憶部２０１に記憶されている標準文書集合解析結果を対象文書集合全体特徴算出部１０５へ転送すると共に、対象文書集合解析結果記憶部２０２に記憶されている対象文書集合解析結果を対象文書集合全体特徴算出部１０５へ転送する。

【００５０】ステップ１０９）　対象文書集合全体特徴算出部１０５は、監視制御部１０から転送された標準文書集合解析結果と対象文書集合解析結果に基づいて対象文書集合全体特徴算出処理を実行し、処理結果を監視制御部１０に出力する。監視制御部１０は、対象文書集合全体特徴算出部１０５から出力された対象文書集合全体特徴を対象文書集合全体特徴記憶部２０３に転送し、その内容を更新し、新たに転送された値を記憶・保持する。さらに、監視制御部１０は、対象文書集合解析結果記憶部２０２に記憶されている対象文書集合解析結果を個別文書特徴算出部１０６へ転送すると共に、個別文書解析結果記憶部２０４に記憶されている個別文書集合解析結果を個別文書特徴算出部１０６へ転送する。

【００５１】ステップ１１０）　個別文書特徴算出部１０６は、転送されてきた対象文書集合解析結果と個別文書解析結果に基づいて個別文書特徴算出処理を実行する。処理結果である、個別文書特徴情報は、監視制御部１０に出力する。監視制御部１０は、個別文書特徴算出部１０６から出力された個別文書特徴情報を個別文書特徴記憶部２０５に転送し、その内容を更新し、新たに転送された値を記憶・保持する。監視制御部１０は、対象文書集合全体特徴記憶部２０３に記憶されている対象文書集合全体特徴を特徴情報算出部１０７へ転送する。それと共に、個別文書特徴記憶部２０５に記憶されている個別文書特徴を特徴情報算出部１０７へ転送する。

【００５２】ステップ１１１）　特徴情報算出部１０７は、対象文書集合全体特徴と個別文書特徴に基づいて特徴情報算出処理を実行し、処理結果である特徴情報を監視制御部１０へ出力する。監視制御部１０は、特徴情報算出部１０７から出力された特徴情報を特徴情報記憶部２０６へ転送し、その内容を更新し、新たに転送された値を記憶・保持する。さらに、監視制御部１０は、特徴

情報記憶部２０６に記憶されている特徴情報を特徴表現抽出部１０８へ転送する。

【００５３】ステップ１１２）　特徴表現抽出部１０８は、転送されてきた特徴情報に基づいて、特徴表現抽出処理を実行し、処理結果である特徴表現を、監視制御部１０へ出力する。監視制御部１０は、特徴表現抽出部１０８から出力された特徴表現を特徴表現記憶部２０７へ転送し、その内容を更新し、新たに転送された値を記憶・保持する。さらに、監視制御部１０は、特徴表現記憶部２０７に記憶されている特徴表現を特徴表現出力部１０９へ転送する。

【００５４】ステップ１１３）　特徴表現出力部１０９は、転送されてきた特徴表現について特徴表現出力処理を実行する。処理結果は、監視制御部１０へ出力され、監視制御部１０において、特徴情報出力部１０９から出力された特徴情報出力処理結果を転送媒体３０へ出力する。

ステップ１１４）　すべての処理が終了か否かを判定し、すべての処理が終了している場合には、当該監視制御処理を終了する。また、終了していない場合には、ステップ１０５に移行し、上述の処理を繰り返す。

【００５５】以下に、各部の詳細な処理について説明する。標準文書集合更新部１０１では、監視制御部１０から転送された標準文書集合に対して標準文書集合更新処理が実行される。この処理は、以降の処理の前処理であり、入力された標準文書集合中の各文書から本装置による処理に必要のない部分を除去する。また、以降の処理で対応している文字コードへ変換される。処理結果は、監視制御部１０へ出力される。

【００５６】標準文書集合解析部１０２では、監視制御部１０から転送される標準文書集合更新処理結果に対して標準文書集合解析処理が実行される。この処理は、文書毎に、その文書に記述されている文章を形態素解析し、各単語の表現及び転送されてきた標準文書集合中での各単語の出現頻度を対応付けて記録するものである。解析結果は、単語表現、出現頻度の２つのカラムからなるテーブルとして監視制御部１０へ出力される。このテーブルにおいて、各行には、各単語表現、及びその単語の出現頻度が記述される。また、このテーブルは、各単語表現をキーとして対応する全ての行を検索できる構造をとる。

【００５７】対象文書集合入力部１０３では、監視制御部１０から転送されてくる対象文書集合に対して対象文書集合入力処理が実行される。この処理は、以降の処理の前処理であり、転送された対象文書集合中の各文書から本装置による処理に必要のない部分を除去する。また、以降の処理で対応している文字コードへ変換される。処理結果は監視制御部１０へ出力される。

【００５８】対象文書集合解析部１０４では、監視制御部１０から転送されてくる対象文書集合入力処理結果に対して対象文書集合解析処理が実行され、処理結果として対象文書集合解析結果、個別文書解析結果が監視制御部１０へ出力される。対象文書集合解析処理は、まず、文書毎に、その文書に記述されている文章を形態素解析し、各単語の表現及び、その文書中での各単語出現頻度を対応付け、文書毎のテーブルに記録する。このテーブルは、対象文書集合中のすべての文書に対して、文書毎に１個作られる。これらのテーブルが、個別文書解析結果として監視制御部１０へ出力される。次に、これらのすべてのテーブルに対して同一単語表現の出現頻度が合算され、対象文書集合全体に対して、１個のテーブルが作られる。このテーブルが対象文書集合解析結果として監視制御部１０へ出力される。上記のすべてのテーブルは、単語表現を記録するカラムと、その単語の出現頻度を記録するカラムの２つのカラムから構成される。また、このテーブルは、各単語表現をキーとして対応するすべての行を検索できる構造をとる。

【００５９】対象文書集合全体特徴算出部１０５では、監視制御部１０から転送されてくる、標準文書集合解析結果と対象文書集合解析結果に基づいて対象文書集合全体特徴を算出する。対象文書集合全体特徴の算出手順は、次のようになる。

①　標準文書集合解析結果を基準情報一時記憶部２０８に転送する。

【００６０】②　対象文書集合解析結果を目的情報一時記憶部２０９に転送する。

③　目的情報特徴スコア算出部１１０を起動し、算出結果が返されるのを待機する。

④　目的情報特徴スコア算出部１１０から算出結果が返されると、それを対象文書集合全体特徴として監視制御部１０へ出力する。

【００６１】個別文書特徴算出部１０６では、監視制御部１０から転送されてくる、対象文書集合解析結果と個別文書解析結果に基づいて、対象文書集合中に含まれている各文書毎に個別文書特徴を算出する。個別文書特徴の算出手順は、以下のようになる。

①　対象文書集合解析結果を基準情報一時記憶部２０８に転送する。

【００６２】②　個別文書解析結果を目的情報一時記憶部２０９に転送する。

③　目的情報特徴スコア算出部１１０を起動し、算出結果が返されるのを待機する。

④　目的情報特徴スコア算出部１１０から算出結果が返されると、それを個別文書特徴として監視制御部１０へ出力する。

【００６３】特徴情報算出部１０７では、監視制御部１０から転送されてくる、対象文書集合特徴と個別文書特徴に基づいて、対象文書集合中の各文書毎に、特徴情報を算出する。算出結果は、監視制御部１０へ出力される。特徴情報は、各単語表現とその単語の特徴を数値化

した特徴情報スコアの２つのカラムからなるテーブルとして表現される。各行は、各単語表現と特徴情報スコアの対応を表す。

【００６４】また、個別文書毎に１つのテーブルが構成される。特徴情報算出の手順は、次のようになる。

① 各単語毎に対象文書集合特徴中での特徴スコアと個別文書特徴中での特徴スコアを読み出す。

② これらのスコアの積を求める。この結果が各単語の特徴情報スコアとなる。

【００６５】対象文書集合中に特徴的単語であり、かつ、個別文書中でも特徴的な単語ほど、この特徴情報スコアは大きな数値を持つ。特徴表現抽出部１０８では、監視制御部１０から転送されてくる特徴情報を用いて、対象文書集合中の各文書から各文書毎にその文書に特徴的な表現（特徴表現）を抽出し、その表現が抽出された文書と対応付け、監視制御部１０へ出力する。

【００６６】各文書の特徴表現とは、各文書に含まれる、予め決められた単語数の連続した単語列、または、予め決められた数の文、あるいは、予め決められた文書中の部分構造（例えば、段落）を構成する単語列であり、その連続する単語列、文、部分構造を構成する単語列に含まれる各単語の特徴情報スコアの平均が最大の部分である。

【００６７】特徴表現として、各文書から一文を抽出する場合、特徴表現抽出の手順は以下のようになる。

① まず、文書中の全単語に特徴情報スコアを付与する。

② 次に、文書中の各文毎に、その文を構成する単語に付与されている特徴情報スコアの平均を求める。

【００６８】③ 特徴情報スコアの平均が最大の文をその文書の特徴表現として抽出する。特徴表現出力部１０９では、監視制御部１０から転送されてくる、各個別文書に対応する特徴表現を監視制御部１０を通して転送媒体３０に出力する。目的情報特徴スコア算出部１１０では、基準情報一時記憶部２０８と目的情報一時記憶部２０９に各々記憶・保存されている基準情報と目的情報に基づいて、特徴スコアを算出し、出力する。

【００６９】特徴スコアとは、基準情報中の単語の出現頻度分布と目的情報中の単語の出現頻度分布を比較し、その分布の相違の大小を各単語毎に数値化したものであり、基準情報中の出現頻度分布と目的情報中の出現頻度分布の相違が大きい単語ほど大きな数値をとる。即ち、目的情報に特徴的な単語ほど、より大きな数値を持つ。

【００７０】特徴スコアは、各単語毎にその単語の出現頻度分布に対して、χ２乗検定の考え方を用いて算出する。χ２乗検定は、「いくつかの群で、ある変数の分布に差があるかどうか」を検定することができる。本発明では、この変数を文書中の単語とする。例えば、標準文書集合に対する対象文書集合の特徴スコアを算出する場合、対象文書集合と標準文書集合中の全単語の出現総数

と対象文書集合中の全単語の出現総数と標準文書集合中の全単語の出現総数から計算される各単語の各文書集合中での出現頻度の期待値の分布と、実際に観測される各単語の各文書集合中での出現頻度の分布からχ２乗値を算出する。この値が大きくなるほど分布に差があることになり、そのような単語ほど偏って出現していることになる。本発明では、この値を用いて各単語の特徴スコアを算出する。

【００７１】以下に標準文書集合と対象文書集合が与えられている場合の具体的な例を用いて説明する。図５は、本発明の一実施例の標準文書集合の一部の例を示し、図６は、本発明の一実施例の対象文書集合の一部の例を示す。以下に各処理部における詳細な処理動作を前述のフローチャートに基づいて説明する。

【００７２】まず、標準文書集合が更新されているか否かが判定される（ステップ１０１）。この例では更新されていると判明したものとする。図５に示す標準文書集合が、標準文書集合更新部１０１に転送される（ステップ１０２）。なお、標準文書集合は、文書集合であり、その文書数は十分に大きいことが望ましい。標準文書集合更新部１０１では、標準文書集合中の各文書から以後の処理に不要である部分が除去される（ステップ１０３）。例えば、ＨＴＭＬ形式の文書の場合は、ＨＴＭＬタグが除去される。また、ワープロ文書の場合は、文字飾り等が除去される。さらに、文書を構成している文字のコードがまちまちである場合は、１つのコードに統一される。この結果は、監視制御部１０に出力される。監視制御部１０は、これを標準文書集合解析部１０２に転送する。

【００７３】標準文書集合解析部１０２は、転送されてきた標準文書集合を解析する（ステップ１０４）。即ち、各文書毎にその文書が記述している文章を形態素解析し、標準文書集合中の単語表現とその単語の出現頻度を求める。図７は、本発明の一実施例の標準文書集合解析結果の例を示す。同図に示す標準文書集合解析結果は、監視制御部１０に出力される。監視制御部１０は、標準文書集合解析結果を標準文書集合解析結果記憶部２０１に転送する。標準文書集合解析結果記憶部２０１は、転送されてきた標準文書集合解析結果を記憶・保持する。

【００７４】以上により、標準文書集合に関する処理が完了する。次に、対象文書集合が入力されているか否かが判定される（ステップ１０５）。入力されている場合は、以下のように処理が進行する。図６に示す対象文書集合が、対象文書集合入力部１０３に入力される（ステップ１０６）。対象文書集合入力部１０３は、入力された文書集合から以降の処理に不要の部分を除去する。また、以降の処理に対応する文字コードへ変換する。処理結果は、監視制御部１０に出力される。監視制御部１０は、対象文書集合入力部１０３から出力された対象文書

集合を対象文書集合解析部１０４に転送する。

【００７５】対象文書集合解析部１０４は、転送された対象文書集合を解析する（ステップ１０７）。解析結果は対象文書集合解析結果と個別文書解析結果である。対象文書集合解析結果は、対象文書集合を構成する単語表現と各単語の出現頻度を記録したテーブルである。この結果の一部を図８に示す。また、個別文書解析結果は、対象文書集合を構成する各文書毎のその文書を構成する単語表現と各単語の出現頻度を記録したテーブルである。この結果の一部を図９に示す。これらの結果は監視制御部１０に出力される。

【００７６】監視制御部１０は、対象文書集合解析結果を対象文書集合解析結果記憶部２０２に、個別文書解析結果を個別文書解析結果記憶部２０４にそれぞれ転送する。対象文書集合解析結果記憶部２０２は、転送されてきた対象文書集合解析結果を記憶・保持する。同様に、個別文書解析結果記憶部２０４は、転送されてきた個別文書解析結果を記憶・保持する（ステップ１０８）。

【００７７】次に、対象文書集合全体特徴を計算する（ステップ１０９）。まず、監視制御部１０は、標準文書集合解析結果記憶部２０１に記憶・保持されている標準文書集合解析結果と、対象文書集合解析結果記憶部２０２に記憶・保持されている対象文書集合解析結果を対象文書集合全体特徴算出部１０５に転送する。対象文書集合全体特徴算出部１０５は、転送されてきた標準文書集合解析結果を基準情報一時記憶部２０８に、同様に、転送されてきた対象文書集合解析結果を目的情報一時記憶部２０９に転送する。

【００７８】次に、特徴スコア算出部１１０を起動する。特徴スコア算出部１１０では、基準情報一時記憶部２０８と目的情報一時記憶部２０９を参照して特徴スコアを算出する。算出結果は、対象文書集合全体特徴算出部１０５に出力させる。図１０は、特徴スコアの算出結果を示す。対象文書集合全体特徴算出部１０５では、特徴スコア算出部１１０の出力結果を対象文書集合全体特徴として図１０に示すような結果を監視制御部１０に出力する。監視制御部１０は、対象文書集合全体特徴を対象文書集合全体特徴記憶部２０４に転送する。対象文書集合全体特徴記憶部２０４は、転送されてきた対象文書全体特徴を記憶・保持する。

【００７９】次に、対象文書集合中の各文書毎に、個別文書特徴を計算する（ステップ１１０）。まず、監視制御部１０は、対象文書集合解析結果記憶部２０２に記憶・保持されている対象文書集合解析結果を個別文書特徴算出部１０６に転送する。また、監視制御部１０は、個別文書解析結果記憶部２０４に記憶・保持されている個別文書解析結果を、一文書分毎に個別文書特徴算出部１０６に転送する。

【００８０】次に、特徴スコア算出部１１０を起動する。特徴スコア算出部１１０では、基準情報一時記憶部

２０８と目的情報一時記憶部２０９を参照して特徴スコアを算出する。算出結果は、個別文書特徴算出部１０６に出力される。その結果の一部を図１１に示す。個別文書特徴算出部１０６では、目的特徴スコア算出部１１０の出力結果を個別文書特徴として監視制御部１０に出力する。その結果一部を図１１に示す。監視制御部１０は、個別文書特徴算出部１０６で得られた個別文書特徴を個別文書特徴記憶部２０５に転送する。個別文書特徴記憶部２０５は、転送されてきた個別文書特徴を記憶・保持する。

【００８１】次に、対象文書集合中の各文書毎に特徴情報を算出する（ステップ１１１）。まず、監視制御部１０は、対象文書集合全体特徴記憶部２０３に記憶・保持されている対象文書集合全体特徴を特徴情報算出部１０７に転送する。また、同様に、個別文書特徴記憶部２０５に記憶・保持されている個別文書特徴を各文書毎に特徴情報算出部１０７に転送する。

【００８２】特徴情報算出部１０７は、監視制御部１０から転送されてきた対象文書全体特徴と個別文書特徴を用いて特徴情報を算出し、算出結果を監視制御部１０へ出力する。算出結果の一部を図１２に示す。ここで、特徴情報とは、各単語毎に対象文書集合中での特徴スコアと個別文書中での特徴スコアを掛けた数値であり、対象文書集合中に特徴的単語であり、かつ、個別文書でも特徴的な単語程大きな数値を持つ。

【００８３】監視制御部１０は、特徴情報算出部１０７の出力結果を各文書毎に特徴情報記憶部２０６に転送する。特徴情報記憶部２０６は、転送されてきた特徴情報を文書毎に記憶・保持する。次に、対象文書集合中の各文書から特徴表現を抽出する（ステップ１１２）。まず、監視制御部１０は、特徴情報記憶部２０６に記憶・保持されている特徴情報を特徴表現抽出部１０８に転送する。

【００８４】特徴表現抽出部１０８では、監視制御部１０から転送されてきた特徴情報を用いて各文書から特徴表現を抽出し、その結果を監視制御部１０へ出力する。抽出結果を図１３に示す。監視制御部１０は、特徴表現抽出部１０８の出力結果を各文書毎に特徴表現記憶部２０７に転送する。

【００８５】特徴表現記憶部２０７は、監視制御部１０から転送されてきた特徴表現を文書毎に記憶・保持すると共に、監視制御部１０に出力する。監視制御部１０は、特徴表現を特徴表現出力部１０９に転送する。監視制御部１０は、対象文書集合中のすべての文書において、特徴表現抽出を終了後、特徴表現出力部１０９は、特徴表現記憶部２０７に記憶・保持されている特徴表現をそれが抽出された文書と対応付けし、転送媒体３０に出力する（ステップ１１３）。

【００８６】以上の実施例において、種々の定義値を用いているが、これらの値は設計値であり、下記のように

必要に応じて変更してもよい。

・特徴情報の算出にχ２乗検定の考え方を用いているが、他の手法で算出してもよい。

・特徴量スコアの算出単位として単語を用いたが、この単位は文字や一定長の文字列でもよい。

【００８７】また、上記の実施例では、図３に示す構成に基づいて説明しているが、この例に限定されることなく、専用のハードウェア回路によって実現することも可能であり、さらに、プログラムされたコンピュータによって実現することも可能である。つまり、監視制御部１０、標準文書集合更新部１０１、標準文書集合解析部１０２、対象文書集合入力部１０３、対象文書集合解析部１０４、対象文書集合全体特徴算出部１０５、個別文書特徴算出部１０６、特徴情報算出部１０７、特徴表現抽出部１０８、特徴表現出力部１０９、目的情報特徴スコア算出部１１０をプログラムとして構築し、文書情報抽出装置として利用されるコンピュータに接続されるディスク装置や、フロッピーディスク、ＣＤ－ＲＯＭ等の可搬記憶媒体に格納しておき、本発明を実施する際に、インストールすることにより、容易に本発明を実現することが可能である。

【００８８】なお、本発明は、上記の実施例に限定されることなく、特許請求の範囲内で種々変更・応用が可能である。

【００８９】

【発明の効果】上述のように、本発明によれば、予め情報抽出知識やパターンなどを用意することなく、文書情報を抽出することが可能となる。これにより、使用開始時に想定したものと対象とする文書内容に差異が生じた場合や、新たな情報を含んでいる場合においても、適切に文書情報を抽出することが可能である。

【００９０】また、文書集合中の各文書を比較するのに適した各文書を特徴付ける文書情報が抽出できるので、文書検索システムの出力編集装置に適用することにより、効率的に検索結果の文書集合から文書を選択、閲覧することができる。

【図面の簡単な説明】

【図１】本発明の原理を説明するための図である。

【図２】本発明の原理構成図である。

【図３】本発明の文書情報抽出装置の構成図である。

【図４】本発明の一実施例の文書情報抽出処理のフローチャートである。

＊【図５】本発明の一実施例の標準文書集合の例である。

【図６】本発明の一実施例の対象文書集合の例である。

【図７】本発明の一実施例の標準文書集合解析結果の例である。

【図８】本発明の一実施例の対象文書集合解析結果の例である。

【図９】本発明の一実施例の個別文書解析結果の例である。

【図１０】本発明の一実施例の対象文書集合全体特徴の例である。

【図１１】本発明の一実施例の個別文書特徴の例である。

【図１２】本発明の一実施例の特徴情報の例である。

【図１３】本発明の一実施例の特徴表現の例である。

【符号の説明】

１　第１の特徴情報算出手段
２　第２の特徴情報算出手段
３　個別文書特徴抽出手段
４　特徴情報出力手段
１０　監視制御部
２０　入力記憶装置
３０　転送媒体
１０１　標準文書集合更新部
１０２　標準文書集合解析部
１０３　対象文書集合入力部
１０４　対象文書集合解析部
１０５　対象文書集合全体特徴算出部
１０６　個別文書特徴算出部
１０７　特徴情報算出部
１０８　特徴表現抽出部
１０９　特徴表現出力部
１１０　目的情報スコア算出部
２０１　標準文書集合解析結果記憶部
２０２　対象文書集合解析結果記憶部
２０３　対象文書集合全体特徴記憶部
２０４　個別文書解析結果記憶部
２０５　個別文書特徴記憶部
２０６　特徴情報記憶部
２０７　特徴表現記憶部
２０８　基準情報一時記憶部
２０９　目的情報一時記憶部

＊

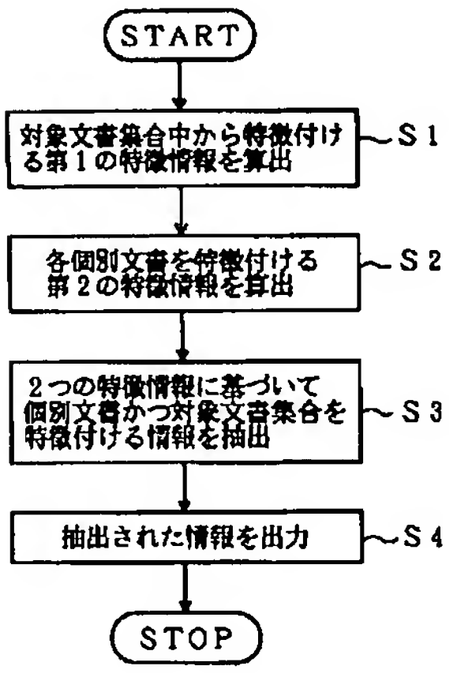
【図１３】

本発明の一実施例の特徴表現の例

千鳥ヶ淵周辺ちどりがふちしゅうへん東京都／千代田区

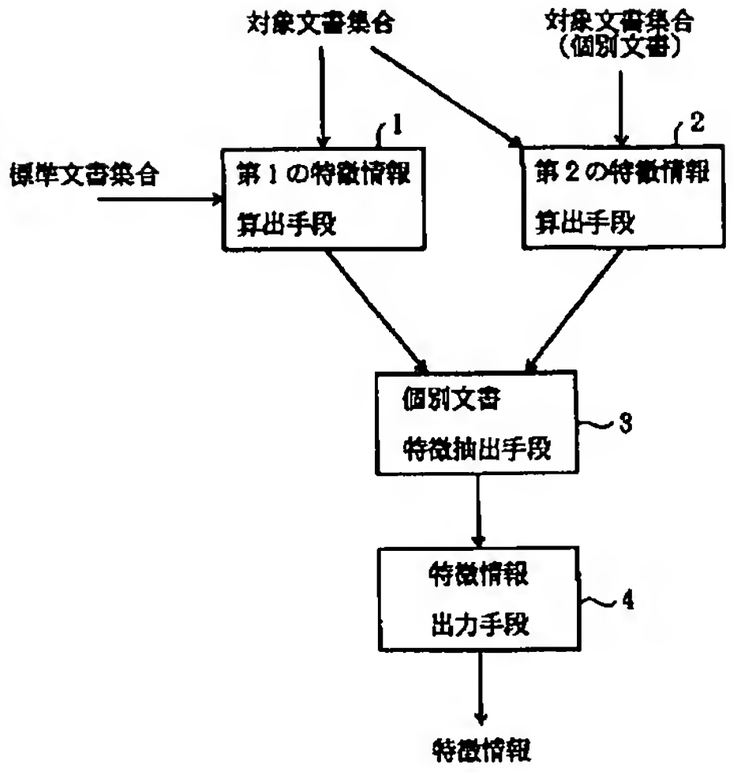
【図１】

本発明の原理を説明するための図



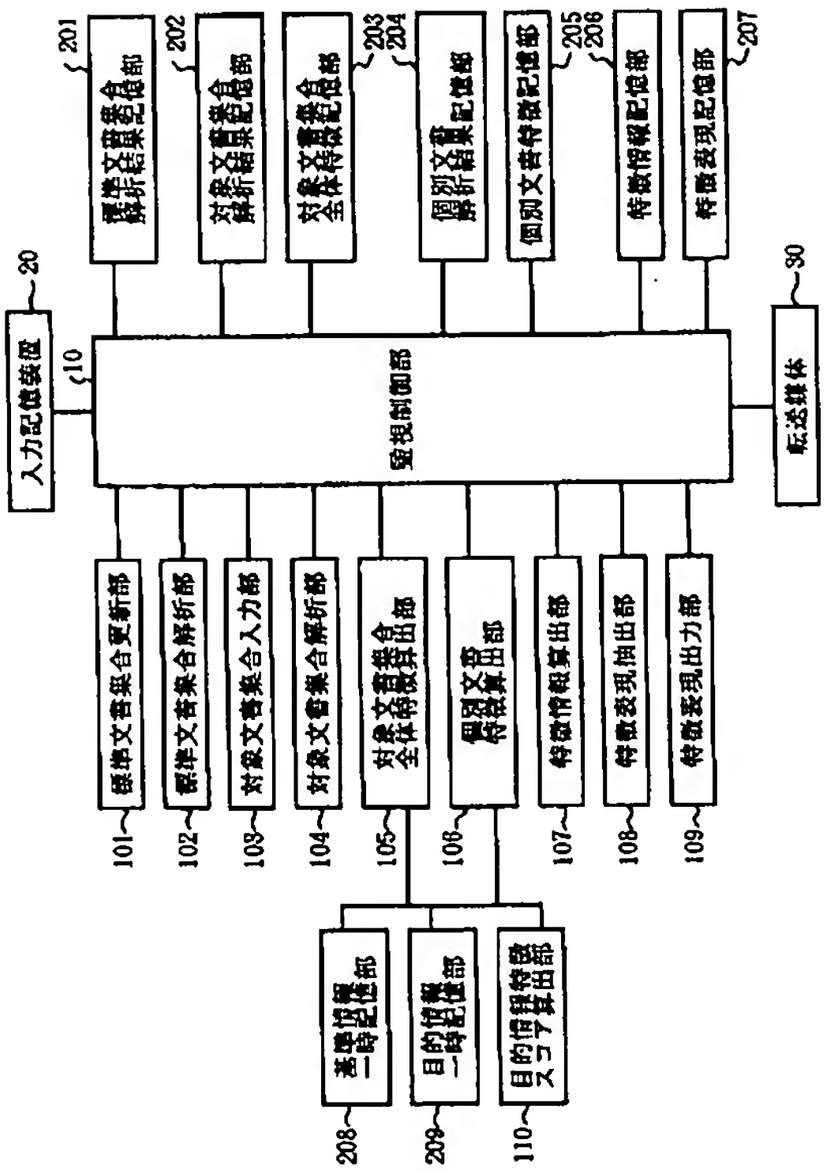
【図２】

本発明の原理構成図



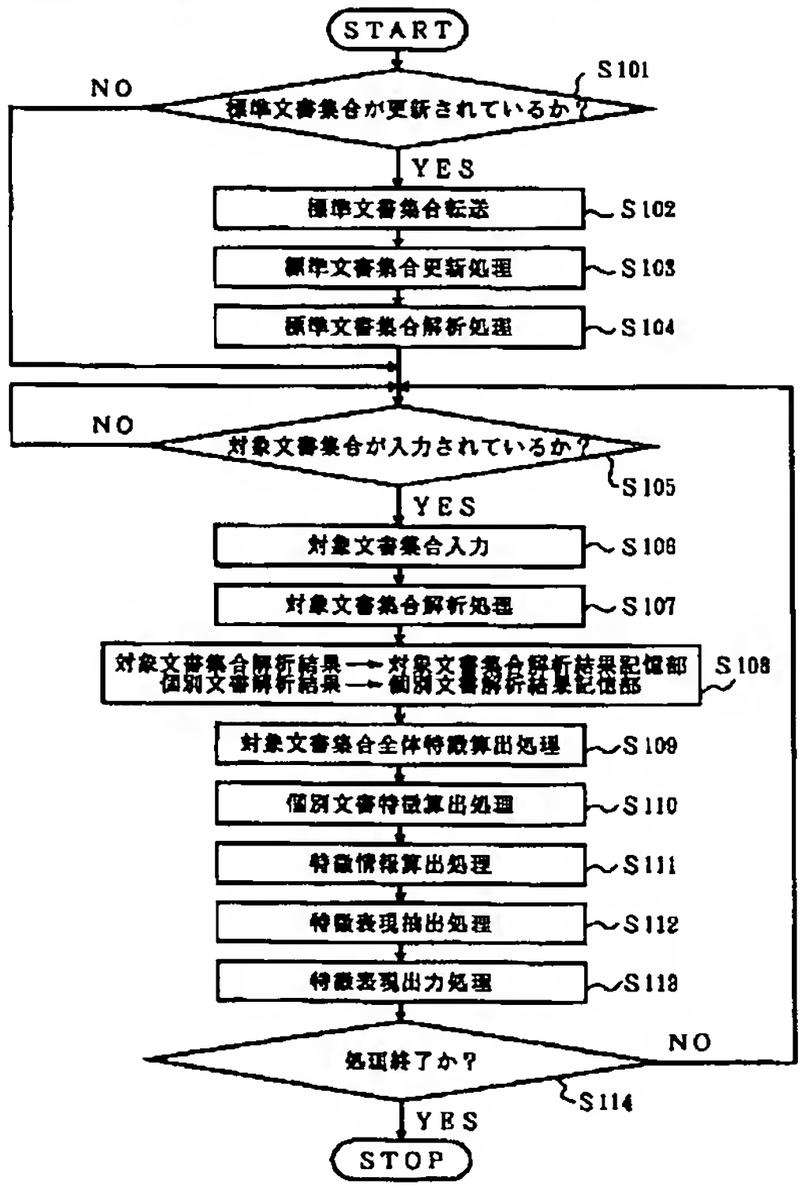
【図３】

本発明の文書情報抽出装置の構成図



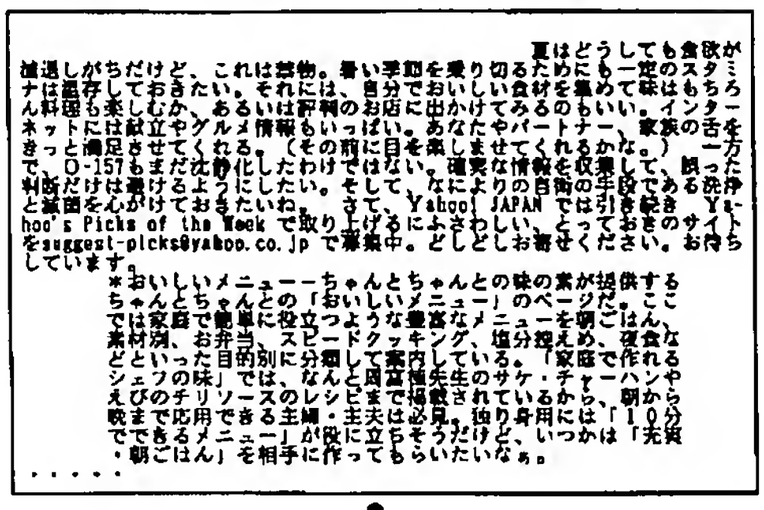
【図４】

本発明の一実施例の文書情報抽出処理のフローチャート



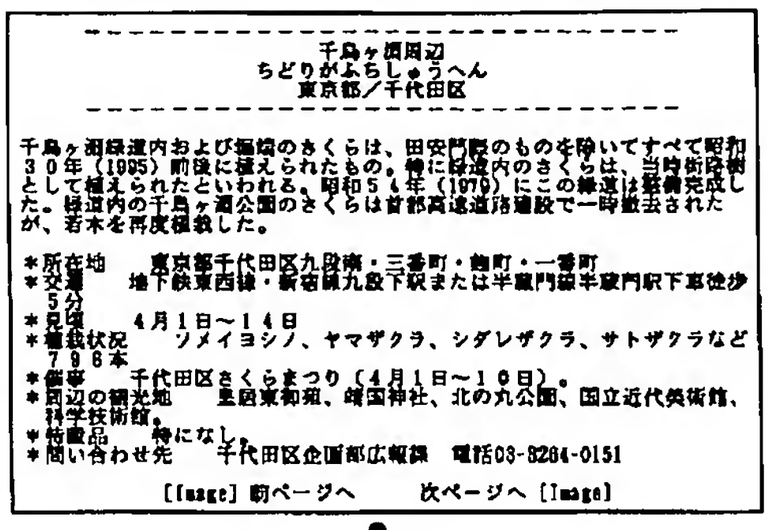
【図５】

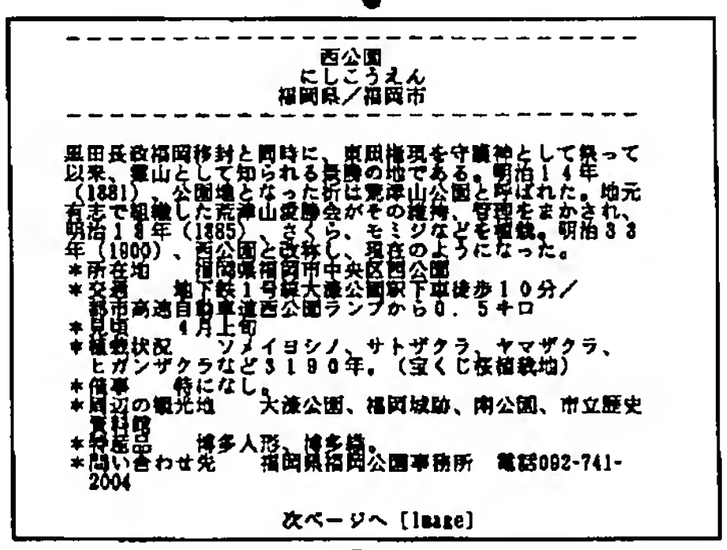
本発明の一実施例の標準文書集合の例



【図６】

本発明の一実施例の対象文書集合の例





【図７】

本発明の一実施例の標準文書集合解析結果の例

単語	頻度	単語	頻度	単語	頻度
ページ	１４	長崎	4	面	2
年	１１	修学旅行	3	用意	2
の	１１	朝	3	評判	2
月	１０	活動	3	販売	2
日	１０	味	3	ちゃん	2
号	9	分	3	洞窟	2
製品	7	意見	3	今	2
皆さん	7	目的	3	食欲	2
こと	7	自分	3	宣伝	2
情報	6	人	3	収納	2
メニュー	6	語	3	技術	2
マーケティング	6	スタッフ	3	ゲスト	2
資金	6	市	3	以外	2
掲載	5	広島	3	彼	2
歌舞伎	5	流通	3	日本語	2
ジャズ	5	アップル	3	中心	2
ため	5	場	3	空手	2
計画	5	段階	2	料理	2
店	5	標準	2	準備	2
インターネット	4	あなた	2	一度	2
支援	4	式典	2	型	2
絵	4	日本	2	別	2
歴史	4	大会	2	インプット	2
紹介	4	選択	2	現在	2
ホーム	4	対応	2	朝	2
部門	4	効果	2	毎月	2
プログラム	4	個々	2	会館	2
浮世絵	4	生態	2	名	2
よう	4	レコーディング	2	確保	2
米	4	お客様	2	方	2

●●●

【図８】

本発明の一実施例の対象文書集合解析結果の例

単語	頻度	単語	頻度	単語	頻度
さくら	5	半蔵門	2	照	1
西公園	4	特産	2	神社	1
千代田区	4	徒歩	2	霊山	1
緑	4	下車	2	守護神	1
緑	4	山	2	整備	1
福岡	3	所在地	2	相撲	1
園	3	状況	2	技術	1
千鳥	3	大濠公園	2	にし	1
周辺	3	町	2	自動車道	1
ページ	3	東	2	現在	1
福岡県	3	もの	2	シダレザクラ	1
駅	3	品	2	広報課	1
4月	3	館	2	博多	1
明治	3	次	2	歴史	1
道内	3	見頃	2	首都	1
公園	3	荒津	2	皇居	1
電話	2	先	2	ヒガンザクラ	1
事	2	改称	1	折	1
昭和	2	上旬	1	ちどり	1
高速	2	道	1	管理	1
ヤマザクラ	2	街路樹	1	資料館	1
ソメイヨシノ	2	北の丸公園	1	勝	1
関市	2	博多駅	1	１０日	1
東京都	2	城跡	1	ふち	1
観光地	2	都市	1	安	1
地	2	完成	1	長政	1
1日	2	前後	1	１４日	1
問い合わせ	2	建設	1	東西	1
地下鉄	2	門	1	公	1
交通	2	愛	1	黒	1

【図９】

本発明の一実施例の個別文書解析結果の例

単語	頻度
さくら	4
千代田区	4
緑	4
緑	3
千鳥	3
遺内	3
濠	3
1日	2
町	2
もの	2
昭和	2
ページ	2
4月	2
周辺	2
駅	2
東京都	2
半蔵門	2
先	1
ちどり	1
電話	1
首都	1
ふち	1
シダレザクラ	1
皇居	1
田	1
広報課	1
１０日	1
神社	1
１４日	1
館	1

単語	頻度
整備	1
ヤマザクラ	1
門	1
当時	1
事	1
ソメイヨシノ	1
分	1
九段下	1
若木	1
高速	1
交通	1
堀	1
品	1
すべて	1
見頃	1
下車	1
道	1
とし	1
靖国	1
東	1
完成	1
所在地	1
企画部	1
隊	1
一時	1
街路樹	1
道路	1
撤去	1
北の丸公園	1
次	1

単語	頻度
東西	1
徒歩	1
特産	1
美術	1
端	1
状況	1
御苑	1
公園	1
観光地	1
しゅうへん	1
麹町	1
問い合わせ	1
学	1
前後	1
建設	1
安	1
国立	1
地下鉄	1
まつり	1
科	1
九段南	1
新宿	1
技術	1
近代	1
再度	1

【図１０】

本発明の一実施例の対象文書集合全体特徴の例

単語	スコア
植	456.643705463183
0	456.643705463183
1	415.130641330166
栽	373.61757719715
さくら	332.104513064133
3	290.591448931116
5	290.591448931116
公園	211.784913941575
ヶ	211.784913941575
4月	207.565320665083
4	207.565320665083
西公園	166.052256532066
8	166.052256532066
9	166.052256532066
玉名	166.052256532066
線	166.052256532066
寺	166.052256532066
緑	166.052256532066
千代田区	166.052256532066
下車	124.53919239905
千鳥	124.53919239905
駅	124.53919239905
ソメイヨシノ	124.53919239905
7	124.53919239905
徒歩	124.53919239905
淵	124.53919239905
明治	124.53919239905
道内	124.53919239905
観光地	124.53919239905
福岡県	124.53919239905

単語	スコア
見頃	124.53919239905
毎日新聞	124.53919239905
岱	124.53919239905
特産	124.53919239905
山	91.9104164998939
市	91.9104164998939
周辺	91.4804036214429
昭和	91.4804036214429
ヤマザクラ	83.0261282660332
大濠公園	83.0261282660332
熊本県	83.0261282660332
高速	83.0261282660332
荒津	83.0261282660332
1日	83.0261282660332
地下鉄	83.0261282660332
2	83.0261282660332
6	83.0261282660332
半蔵門	83.0261282660332
自動車道	83.0261282660332
岡市	83.0261282660332
蛇	83.0261282660332
町	83.0261282660332
折	83.0261282660332
サトザクラ	83.0261282660332
事	72.3427864813708
問い合わせ	72.3427864813708
状況	72.3427864813708
所在地	72.3427864813708
群	72.3427864813708
交通	59.305729403164

単語	スコア
上旬	54.0254484448309
先	50.0004282908523
福	50.0004282908523
.	41.5130641330166
勢	41.5130641330166
照	41.5130641330166
ノリ	41.5130641330166
樹	41.5130641330166
堀	41.5130641330166
立願	41.5130641330166
滝	41.5130641330166
1955	41.5130641330166
首都	41.5130641330166
名駅	41.5130641330166
勝	41.5130641330166
田	41.5130641330166
山一	41.5130641330166
1960	41.5130641330166
坊	41.5130641330166
黒田	41.5130641330166
公	41.5130641330166
麹町	41.5130641330166
人形	41.5130641330166
1900	41.5130641330166
門	41.5130641330166
前後	41.5130641330166
端	41.5130641330166
完成	41.5130641330166
092-741-2004	41.5130641330166
玉	41.5130641330166

【図１１】

本発明の一実施例の個別文書特徴の例

単語	スコア
緑	8.85496183206107
千代田区	8.85496183206107
道内	6.6412213740458
千鳥	6.6412213740458
淵	6.6412213740458
町	4.42748091603054
もの	4.42748091603054
東京都	4.42748091603054
１日	4.42748091603054
半蔵門	4.42748091603054
線	4.08660708744492
駅	2.39380150814102
堀	2.21374045801527
東西	2.21374045801527
1955	2.21374045801527
若木	2.21374045801527
館	2.21374045801527
国立	2.21374045801527
再度	2.21374045801527
完成	2.21374045801527
科学	2.21374045801527
広報課	2.21374045801527
ふち	2.21374045801527
近代	2.21374045801527
皇居	2.21374045801527
シダレザクラ	2.21374045801527
１０日	2.21374045801527
首都	2.21374045801527
１４日	2.21374045801527
すべて	2.21374045801527

単語	スコア
田	2.21374045801527
美術館	2.21374045801527
新宿	2.21374045801527
建設	2.21374045801527
しゅうへん	2.21374045801527
端	2.21374045801527
神社	2.21374045801527
まつり	2.21374045801527
整備	2.21374045801527
九段南	2.21374045801527
一時	2.21374045801527
麹町	2.21374045801527
前後	2.21374045801527
門	2.21374045801527
九段下	2.21374045801527
当時	2.21374045801527
北の丸公園	2.21374045801527
御苑	2.21374045801527
三	2.21374045801527
技術	2.21374045801527
1979	2.21374045801527
街路樹	2.21374045801527
撤去	2.21374045801527
ちどり	2.21374045801527
03-3264-0151	2.21374045801527
靖国	2.21374045801527
道	2.21374045801527
道路	2.21374045801527
際	2.21374045801527
企画部	2.21374045801527

単語	スコア
ヶ	2.0025074811164
さくら	1.6914042887972
昭和	1.33500498741094
周辺	0.69967103588392
ページ	0.69967103588392
安	0.667502493705468
東	0.596120943710052
ヤマザクラ	0.596120943710052
地下鉄	0.596120943710052
サトザクラ	0.596120943710052
高速	0.596120943710052
次	0.596120943710052
4月	0.271930638619556
品	0.0117976692737065
ソメイヨシノ	0.0117976692737065
車	0.0117976692737065
先	0.0117976692737065
下車	0.0117976692737065
観光地	0.0117976692737065
所在地	0.0117976692737065
交通	0.0117976692737065
電話	0.0117976692737065
見頃	0.0117976692737065
毎日新聞	0.0117976692737065
問い合わせ	0.0117976692737065
状況	0.0117976692737065
徒歩	0.0117976692737065
特産	0.0117976692737065
7	0.00959380705870577

【図１２】

本発明の一実施例の特徴情報の例

単語	スコア
千代田区	1470.38639371906
緑	1470.38639371906
濵	827.092346466973
道内	827.092346466973
千鳥	827.092346466973
線	678.590328430163
さくら	561.722997725584
ヶ	424.100874555597
町	367.596598429766
１日	367.596598429766
半蔵門	367.596598429766
駅	298.122106587511
東京都	136.60466005689
昭和	122.138810129879
1955	91.8991496074414
広報課	91.8991496074414
しゅうへん	91.8991496074414
堀	91.8991496074414
東西	91.8991496074414
ちどり	91.8991496074414
皇居	91.8991496074414
シダレザクラ	91.8991496074414
国立	91.8991496074414
首都	91.8991496074414
１０日	91.8991496074414
若木	91.8991496074414
ふち	91.8991496074414
田	91.8991496074414
再度	91.8991496074414
企画部	91.8991496074414

単語	スコア
神社	91.8991496074414
１４日	91.8991496074414
近代	91.8991496074414
前後	91.8991496074414
靖国	91.8991496074414
門	91.8991496074414
新宿	91.8991496074414
当時	91.8991496074414
九段南	91.8991496074414
まつり	91.8991496074414
九段下	91.8991496074414
建設	91.8991496074414
完成	91.8991496074414
麹町	91.8991496074414
御苑	91.8991496074414
03-3264-0151	91.8991496074414
端	91.8991496074414
北の丸公園	91.8991496074414
撤去	91.8991496074414
1979	91.8991496074414
道路	91.8991496074414
街路樹	91.8991496074414
周辺	64.012485804217
４月	56.4433702037289
ヤマザクラ	49.4936139345395
地下鉄	49.4936139345395
高速	49.4936139345395
サトザクラ	49.4936139345395
館	43.7624975239632
整備	43.7624975239632

単語	スコア
一時	43.7624975239632
美術館	43.7624975239632
道	27.7524977338144
際	27.7524977338144
東	23.5689249194848
もの	20.8193465069299
安	20.5949958834634
すべて	15.008489529514
三	15.008489529514
次	12.489114997965
科学	4.1374457224585
ページ	1.48577944532577
見頃	1.46927220353849
ソメイヨシノ	1.46927220353849
徒歩	1.46927220353849
下車	1.46927220353849
観光地	1.46927220353849
毎日新聞	1.46927220353849
特産	1.46927220353849
7	1.19480498312352
問い合わせ	0.853476269245578
状況	0.853476269245578
所在地	0.853476269245578
事	0.853476269245578
交通	0.69966938153446
技術	0.654223559029463
先	0.589888516519153
品	0.350883341438889
電話	0.286798125124254
5	-8.92323790311528